The Spanish Journal of Psychology
2009, Vol. 12, No. 1, 288-307


ISSN 1138-7416

# Statistical Inference Involving Binomial and Negative Binomial Parameters

Miguel A. García-Pérez[1] and Vicente Núñez-Antón[2]
[1]Universidad Complutense (Spain)
[2]Universidad del País Vasco (Spain)

Statistical inference about two binomial parameters implies that they are both estimated by binomial sampling. There are occasions in which one aims at testing the equality of two binomial parameters before and after the occurrence of the first success along a sequence of Bernoulli trials. In these cases, the binomial parameter before the first success is estimated by negative binomial sampling whereas that after the first success is estimated by binomial sampling, and both estimates are related. This paper derives statistical tools to test two hypotheses, namely, that both binomial parameters equal some specified value and that both parameters are equal though unknown. Simulation studies are used to show that in small samples both tests are accurate in keeping the nominal Type-I error rates, and also to determine sample size requirements to detect large, medium, and small effects with adequate power. Additional simulations also show that the tests are sufficiently robust to certain violations of their assumptions.
*Keywords: statistical tests, binomial parameters, power, effect size, Monte Carlo simulations*

El contraste de hipótesis acerca de dos proporciones supone que cada una de ellas se ha estimado mediante muestreo binomial, pero hay ocasiones en que interesa evaluar la hipótesis de que la probabilidad de éxito a medida que se repite una determinada tarea varía una vez que se ha obtenido el primer éxito. En estos casos, la probabilidad de éxito antes de que ocurra el primer éxito se estima mediante muestreo binomial negativo, en tanto que la probabilidad de éxito después del primer éxito se estima mediante muestreo binomial, y ambas estimaciones están relacionadas. En este trabajo se presentan procedimientos para contrastar dos hipótesis aplicables a esta situación. Una es la de que las dos probabilidades son iguales y tienen un determinado valor; la otra es más general y sólo expresa que las dos probabilidades son iguales. El comportamiento de estos dos contrastes en muestras finitas se analiza mediante simulaciones cuyos resultados muestran que en ambos casos se preserva adecuadamente la tasa nominal de error de tipo I. También se ha determinado mediante simulación los tamaños muestrales necesarios para detectar efectos grandes, medianos o pequeños con potencia suficiente. Finalmente, otro grupo de simulaciones muestra que ambos contrastes son suficientemente robustos ante violaciones de sus supuestos.
*Palabras clave: contraste de hipótesis, proporciones, potencia, tamaño del efecto, simulación*

Acknowledgements: This research was supported by grants SEJ2005-00485 (Ministerio de Educación y Ciencia), MTM2004-00341 (Ministerio de Educación y Ciencia and FEDER), MTM2007-60112 (Ministerio de Educación y Ciencia and FEDER), and IT-334-07 (Departamento de Educación del Gobierno Vasco – UPV/EHU Econometrics Research Group).

Corresponding Author's Address: Miguel A. García-Pérez, Departamento de Metodología, Facultad de Psicología, Universidad Complutense, Campus de Somosaguas, 28223 Madrid, Spain. Phone: +34 913 943 061; Fax: +34 913 943 189; E-mail: miguel@psi.ucm.es

How to cite the authors of this article: García-Pérez, M.A., Núñez-Antón, V.

Consider the hypothesis that the first success in a task alters the subsequent probability of success in the same task. This hypothesis underlies studies of avoidance learning, where the very existence of learning implies that the probability of observing the behavior that is to be avoided decreases after administration of negative reinforcers. This is also the hypothesis underlying priming in studies on inattentional blindness (Neisser & Becklen, 1975), where it is assumed that the probability of detecting an unexpected event increases once the first such event has been detected.

Testing this type of hypothesis involves estimating and comparing binomial parameters before and after the first occurrence of a success in a sequence of Bernoulli trials. By the nature of the process, the binomial parameter $p_a$ before the first occurrence of a success must be estimated under geometric sampling, that is, through the number of Bernoulli trials before the first success; on the other hand, the binomial parameter $p_b$ after the first success is estimated under binomial sampling, that is, through the number of successes in a fixed number of Bernoulli trials. Figure 1 shows sample data for such situation, which will be useful to introduce some notation and terminology. Typically, the data consist of $R$ replicates of a sequence of $T$ Bernoulli trials, each replicate $r$ ($1 \le r \le R$) being split into two parts at the trial where the first success is observed, with the characteristic that the first success occurs on trial $T - 1$ at the latest or otherwise the replicate is discarded. The first part thus consists of a sequence of $y_r^{(a)}$ failures (the random variable in this part) followed by $x_r^{(a)} = 1$ successes for a total of $n_r^{(a)} = y_r^{(a)} + 1$ trials; the second part consists of $n_r^{(b)} = T - n_r^{(a)}$ trials among which $x_r^{(b)}$ (the random variable in this part) turn out to be successes and $y_r^{(b)} = n_r^{(b)} - x_r^{(b)}$ are failures. Thus, $y_r^{(a)}$ has a geometric distribution with parameter $p_a$ (assumed constant across trials), whereas $x_r^{(b)}$ has a (conditional) binomial distribution with parameters $n_r^{(b)}$ and $p_b$ (also assumed constant across trials). If the probabilities $p_a$ and $p_b$ remain invariant across replicates, $y_a = \sum_{r=1}^{R} y_r^{(a)}$ has a negative binomial distribution with parameters $x_a = \sum_{r=1}^{R} x_r^{(a)} = R$ and $p_a$, whereas $x_b = \sum_{r=1}^{R} x_r^{(b)}$ has a (conditional) binomial distribution with parameters $n_b = \sum_{r=1}^{R} n_r^{(b)}$ and $p_b$. Thus, $\hat{p}_a = x_a/n_a$ with $n_a = x_a + y_a$ is the maximum likelihood estimate of $p_a$, whereas $\hat{p}_b = x_b/n_b$ is the maximum likelihood estimate of $p_b$.

A less formal description, also based on Figure 1, follows. Note that (i) $y_a$ is the total number of failures in the first sequence for all replicates (i.e., the total number of zeros on the left of the thick line running down the data array in Figure 1), (ii) $x_a$ is the total number of successes in the first sequence for all replicates (i.e., the total number of ones on the left of the thick line in Figure 1, which makes $x_a = R$ because by definition there is only one success per replicate), (iii) $\hat{p}_a$ is thus the proportion of successes in the first sequence for all replicates (i.e., the proportion of successes on the left of the thick line in Figure 1), and (iv)

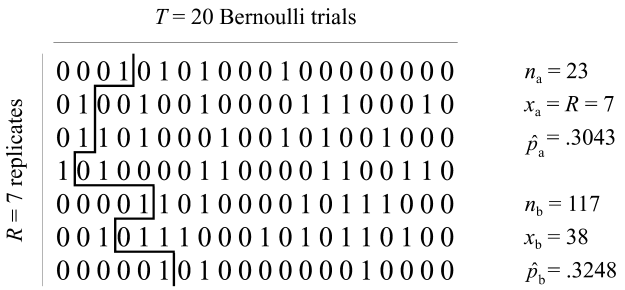


*Figure 1.* Sample data illustrating the situation for a test of equality of negative binomial and binomial parameters. Each cell in the array indicates success (1) or failure (0) at each of the Bernoulli trials in each replicate. The negative binomial case arises for data up to and including the first success in each replicate, and is represented by Bernoulli variates on the left of the thick line running down the data array; the binomial case arises for data after the first success in each replicate, and is represented by variates on the right of the thick line. Numerical values on the right of the data array indicate the numbers of trials ($n$), numbers of successes ($x$), and estimates of the probability of success ($\hat{p}$) in parts a (negative binomial) and b (binomial).

$y_b$ and $x_b$ are the numbers of failures and successes analogously defined for the second sequence across all replicates (i.e., on the right of the thick line in Figure 1) so that $\hat{p}_b$ is the proportion of successes in the second sequence for all replicates.

The practical problem that arises in a case like this is that typical statistical tests for the comparison of two binomial parameters assume that sample proportions are unrelated and estimated under binomial sampling. Under the conditions established in the preceding paragraph, the sample proportion $\hat{p}_a$ comes from a truncated negative binomial distribution for $y_a$ (since there is a maximal number of $T - 1$ trials for each of its constituent geometric variates), whereas the conditional binomial distribution of $x_b$ yields the sample proportion $\hat{p}_b$. This paper proposes test statistics for the comparison of two binomial parameters that come from related samples in the situation just described, that is, in a series of Bernoulli trials that are split into two subsets at the occurrence of the first success in the sequence so that one of the binomial parameters is estimated under negative binomial sampling whereas the other is estimated under binomial sampling. This paper lies out the derivation of the asymptotic distribution of the test statistic in two different cases, presents the results of a series of simulation studies that investigate the small-sample accuracy of the tests and their small-sample power, presents the results of additional simulation studies investigating the robustness of the tests to violation of the assumptions of invariance of $p_a$ and $p_b$ across replicates and invariance of $p_b$ during the entire second stage, and provides an example with empirical data.

https://doi.org/10.1017/S1138741600001694 Published online by Cambridge University Press

## Derivation of the Test Statistic

Consider a fixed number $R$ of replicates of a truncated geometric experiment, each replicate $r$ ($1 \le r \le R$) consisting of a maximum of $T - 1$ Bernoulli trials (with success probability $p_a$) until $x_r^{(a)} = 1$ successes are observed, and with the entire replicate discarded if a success has not occurred within the $T - 1$ trials. Let $y_r^{(a)}$ be the applicable geometric variable and let $n_r^{(a)} = x_r^{(a)} + y_r^{(a)} = 1 + y_r^{(a)}$. Thus, $y_r^{(a)}$ is the number of failures until the first success and $n_r^{(a)}$ is the total number of Bernoulli trials up to and including the first success. Across the $R$ replicates, $y_a = \sum_{r=1}^{R} y_r^{(a)}$ is a negative binomial random variable with parameters $x_a = \sum_{r=1}^{R} x_r^{(a)} = R$ and $p_a$, and let $n_a = \sum_{r=1}^{R} n_r^{(a)} = x_a + y_a$. Then, by the central limit theorem,

$$\frac{1}{\hat{p}_a} = \frac{n_a}{x_a} = \frac{x_a + y_a}{x_a} = 1 + \frac{\sum_{r=1}^{R} y_r^{(a)}}{R} \tag{1}$$

is asymptotically $N\left(\frac{1}{p_a}, \frac{1 - p_a}{R\ p_a^2}\right)$ and, hence,

$$\frac{1 / \hat{p}_a - 1/p_a}{\sqrt{(1 - p_a) / R\ p_a^2}} \text{ is asymptotically N(0, 1).}$$

Consider also a fixed number $R$ of replicates of a binomial experiment, where replicate $r$ ($1 \le r \le R$) moves ahead the sequence of trials in the $r$-th negative binomial experiment described above to an overall total of $T$ trials so that the binomial experiment $r$ has parameters $n_r^{(b)} = T - n_r^{(a)}$ and $p_b$. Let $x_r^{(b)}$ be the applicable binomial variable in replicate $r$ and let $y_r^{(b)} = n_r^{(b)} - x_r^{(b)}$. Across the $R$ replicates, $x_b = \sum_{r=1}^{R} x_r^{(b)}$ is a binomial random variable with parameters $n_b = \sum_{r=1}^{R} n_r^{(b)}$ and $p_b$, and let $y_b = \sum_{r=1}^{R} y_r^{(b)} = n_b - x_b$. Then, by the central limit theorem,

$$\hat{p}_b = \frac{x_b}{n_b} = \frac{\sum_{r=1}^{R} x_r^{(b)}}{n_b} \tag{2}$$

is asymptotically $N\left(p_b, \frac{p_b(1 - p_b)}{n_b}\right)$ and, hence,

$$\frac{\hat{p}_b - p_b}{\sqrt{p_b\,(1 - p_b) / n_b}} \text{ is asymptotically N(0, 1).}$$

Now, testing the null hypothesis $H_0$: $p_a = p_b$ is equivalent to testing $H_0$: $p_b/p_a = 1$. This can be done with a test statistic based on the ratio $\hat{p}_b / \hat{p}_a$ or a transformation thereof. Under the null hypothesis of identity, there is a single underlying probability $p = p_a = p_b$ and it can then be shown (see Appendix A) that the statistic

$$B = \frac{\hat{p}_b - p_b}{\sqrt{p(1 - p)/n_b}} \quad \frac{1/\hat{p}_a - 1/p}{\sqrt{(1 - p)/R\,p^2}} =$$

$$= \frac{(\hat{p}_b - p)\ (1/\hat{p}_a - 1/p)\ \sqrt{p\ n_b\ R}}{1 - p} \tag{3}$$

—which is the product of the two N(0, 1) variables introduced above—is asymptotically distributed as

$$f(b) = \frac{K_0\,(|b|)}{\pi} = \frac{1}{\pi} \int_0^{\infty} \frac{\cos\,(|b|t)}{\sqrt{1 + t^2}}\,\mathrm{d}t, \tag{4}$$

where $K_0$ is the modified Bessel function of the second kind and zero order. This function is available in subroutine packages such as the so-called numerical recipes (Press, Flannery, Teukolsky, & Vetterling, 1986), IMSL (Visual Numerics, Inc., 1997), or NAG (Numerical Algorithms Group, 1999) and it is also available in software packages such as Mathematica (Wolfram, 1992) or MATLAB (The MathWorks, Inc., 2004).

Since $B \sim K_0(|b|)/\pi$, it follows that $B' = \hat{p}_b / \hat{p}_a = \gamma B - \delta$ with

$$\gamma = (1 - p)/\sqrt{p\ n_b R}, \tag{5}$$

$$\delta = 1 - \hat{p}_b / p - p / \hat{p}_a \tag{6}$$

is distributed as $K_0(|b' + \delta|/\gamma)/\pi\gamma$. Note that parameters $\gamma$ and $\delta$ are functions of (i) the fixed quantity $x_a$, which equals the number $R$ of replicates; (ii) the random variable $n_b = RT - n_a = R(T - 1) - y_a$, which is a linear transformation of the random variable $y_a$ distributed as negative binomial with parameters $R$ and $p$ and ranging from $R$ to $R(T - 1)$; and (iii) the unknown population parameter $p$. Next we consider two separate cases which vary as to the status of the population parameter $p$.

### *Case 1. Testing $H_0$: $p_b = p_a = p_0$.*

This hypothesis tests for a given common value $p_0$ for both binomial parameters, yielding an essentially two-tailed test. Note also that the hypothesis assumes that both binomial parameters are indeed equal and that we do not aim at testing their equality *per se*. Testing this hypothesis amounts to computing $B' = \hat{p}_b / \hat{p}_a$ from the data and comparing its value with the critical limits $b'_{\alpha/2}$ and $b'_{1-\alpha/2}$ obtained from the scaled Bessel distribution for a two-sided size-$\alpha$ test using $\hat{p}_a$, $\hat{p}_b$, and $n_b$ from the data and $p = p_0$ from the null hypothesis. From the linear transformation relating $B'$ to $B$, it follows straightforwardly that $b'_\nu = \gamma\, b_\nu - \delta$, where $b_\nu$ is the critical limit from the Bessel distribution in Equation (4). Although these critical limits are easily obtained with the software mentioned above, Appendix B tabulates $b_\nu$ for $\nu \in [.9, .999]$.

https://doi.org/10.1017/S1138741600001694 Published online by Cambridge University Press

*Case 2. Testing $H_0$: $p_b / p_a = 1$.*

This case aims at testing for equality of the two binomial parameters, whatever their value may be. Then, the occurrence of $p$ in Equations (5) and (6) should be replaced with its maximum likelihood estimate under the null hypothesis (as is done in the typical two-sample binomial test), which is given by (see Appendix C)

$$\hat{p}_\bullet = \frac{x_a + x_b}{n_a + n_b} = \frac{R + x_b}{RT} \tag{7}$$

The distribution of the test statistic $B' = \hat{p}_b / \hat{p}_a$ in these conditions can be determined by noting that parameter $n_b$ in the binomial distribution during the second stage is actually a random variable whose value depends on the outcomes of the first stage of negative binomial sampling given the unknown value of parameter $p$. Thus, we need to condition on the actual value of $n_b$ and marginalize across values of $x_b$. The joint distribution $f(b', n_b, x_b)$ can be written out as $f(b' \mid x_b, n_b) \times f(x_b \mid n_b) \times f(n_b)$. Given that $n_b$ ranges from $R$ to $R(T-1)$ and has a negative binomial distribution with parameters $R$ and $\hat{p}_\bullet$ and that $x_b$ ranges from 0 to $n_b$ and has a binomial distribution with parameters $n_b$ and $\hat{p}_\bullet$, conditioning on $n_b$ and marginalizing across $x_b$ yields

$$\tilde{f}(b' \mid n_b) = \frac{1}{S} \sum_{x_b=0}^{n_b} \frac{K_0(|b' + \delta| / \gamma)}{\pi\gamma} \binom{n_b}{x_b} \hat{p}_\bullet^{x_b} (1 - \hat{p}_\bullet)^{n_b - x_b} \times \binom{RT - n_b - 1}{R - 1} \hat{p}_\bullet^{R} (1 - \hat{p}_\bullet)^{R(T-1) - n_b} \tag{8}$$

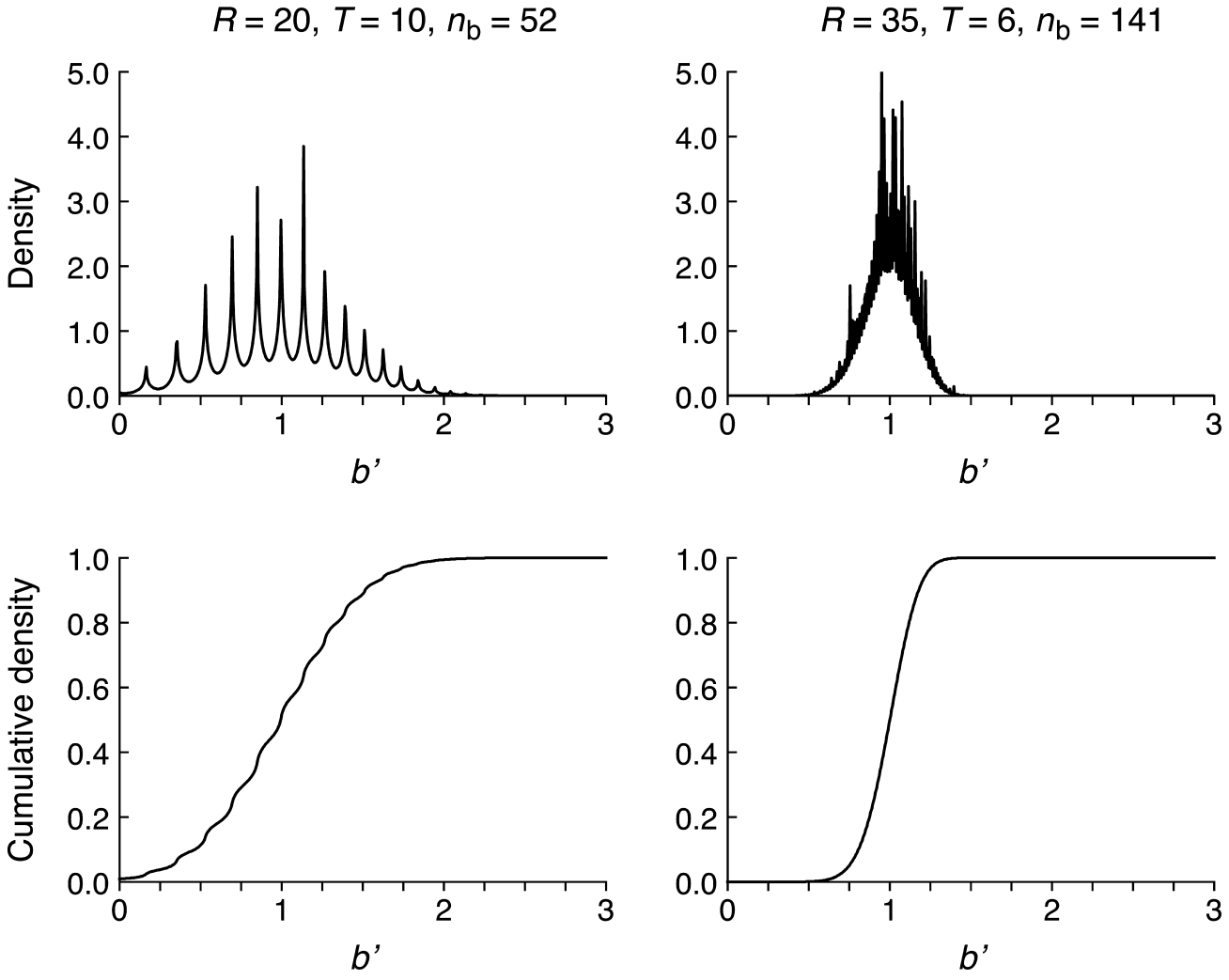


*Figure 2.* Probability density (top panel) and cumulative density (bottom panel) for two sample cases (columns) with values for $R$, $T$, and $n_b$ as given at the top of each column. The spikes in the top panels are each a discontinuity similar to that in the left panel of Figure 13, but these cannot be adequately rendered graphically at the scale of these plots.

https://doi.org/10.1017/S1138741600001694 Published online by Cambridge University Press

with γ and δ defined as in Equations (5) and (6) and $p$ as in Equation (7), all of them with values entirely determined by the dummy variable in the sum, that is,

$$\gamma = \left(1 - \frac{R + x_b}{RT}\right) \Big/ \sqrt{\frac{n_b\,(R + x_b)}{T}}, \tag{9}$$

$$\delta = 1 - \frac{x_b RT}{n_b\,(R + x_b)} - \frac{(R + x_b)\,(RT - n_b)}{R^2 T}, \tag{10}$$

and where

$$S = \sum_{x_b=0}^{n_b} \binom{n_b}{x_b} \hat{p}_\bullet^{x_b}(1 - \hat{p}_\bullet)^{n_b - x_b} \binom{RT - n_b - 1}{R - 1} \times$$

$$\hat{p}_\bullet^{R}(1 - \hat{p}_\bullet)^{R(T-1)-n_b} \tag{11}$$

to rescale the conditional distribution so as to make it a proper probability density function. The density in Equation (8) is a weighted sum of scaled Bessel distributions that is difficult to integrate, but the cumulative distribution is easy to obtain directly through

$$\tilde{F}\,(b' \,|\, n_b) = \frac{1}{S} \sum_{x_b=0}^{n_b} F^*\left(\frac{|\,b' + \delta\,|}{\gamma}\right) \binom{n_b}{x_b} \hat{p}_\bullet^{x_b}(1 - \hat{p}_\bullet)^{n_b - x_b} \times$$

$$\binom{RT - n_b - 1}{R - 1} \hat{p}_\bullet^{R}(1 - \hat{p}_\bullet)^{R(T-1)-n_b} \tag{12}$$

where $F^*$ is the cumulative Bessel distribution in Equation (A5) in Appendix A. Figure 2 shows the probability and cumulative distributions in Equations (8) and (12) for sample values of $R$, $T$, and $n_b$.

Note that, by marginalization and conditioning, this distribution is a function of $T$ and $R$ (which are fixed parameters for each data set) but also of $n_b$ (which has a fixed value after the data have been collected). Determining critical limits from this distribution is simple with the software described above, and a FORTRAN program that computes these limits as well as general $p$-values for empirical $B'$ given $T$, $R$, and $n_b$ is available from the first author. Testing this hypothesis (which accommodates two-tailed and one-tailed tests) amounts to computing $B' = \hat{p}_b / \hat{p}_a$ from the data and comparing its value with the critical limits for a size-α test obtained from the distribution in Equation (12) or the program just mentioned.

## Small-Sample Accuracy

The small-sample accuracy of the test statistic in each of the two cases described in the preceding section was determined using simulation methods that generated data under the null hypothesis of identity of the negative binomial and the binomial parameters. Distinct simulation conditions were defined as the factorial combination of several values for parameter $p$ (between .1 and .9 in steps of .1), several numbers $T$ of trials within each replicate (between 5 and 100 in steps of 5), and several numbers $R$ of replicates (between 5 and 100, in steps of 5). Given the values of $p$, $T$, and $R$ in the current simulation condition, data were generated to fill up a table similar to that in Figure 1 but whose size was adjusted to the current values of $T$ and $R$. The outcome of each of the $T \times R$ Bernoulli trials (all of which have success probability $p$) was simulated using NAG subroutine G05DZF (Numerical Algorithms Group, 1999), and 20,000 such tables were generated per simulation condition. In each of these tables, where $x_a = R$ by definition,[1] $n_a$, $\hat{p}_a$, $x_b$, $n_b$, and $\hat{p}_b$ were computed as illustrated in Figure 1 and the value of $B'$ was computed from these quantities.

The resultant 20,000 values of the test statistic were used to determine empirical Type-I error rates slightly differently in each case. In Case 1, Type-I error rates of two-tailed tests were determined at significance levels $1 - \alpha = .90$, .95, and .99 by computing the proportion of occasions (across the set of 20,000 tables) in which the statistic value exceeded either of the critical limits $b'_{\alpha/2}$ and $b'_{1-\alpha/2}$ for a two-sided size-α test, where these critical limits were determined for each table by obtaining coefficients γ and δ through Equations (5) and (6) using the data from the current table and then transforming the limits in Appendix B through these coefficients. In Case 2, Type-I error rates were determined for two-tailed and one-tailed tests at analogous significance levels by computing the proportion of occasions (across the set of 20,000 tables) in which the statistic value exceeded the applicable critical limit that was obtained numerically from Equation (12).

### *Small-Sample Accuracy in Case 1*

Figure 3 summarizes empirical Type-I error rates at nominal test sizes $\alpha = .1$, .05, and .01 for all of our simulation conditions. Quite clearly, the test is not accurate in all conditions but its failures are easy to describe. To a first approximation, the number $R$ of replicates does not greatly affect accuracy: The 20 curves (one for each value of $R$) making up each of the three bundles in each panel of

[1] If no success was observed across the $T$ trials in a given replicate but also if the first success occurred on trial $T$ (something that was not uncommon when $p$ and $T$ were both small), the entire set of outcomes (either a string of $T$ failures or a string of $T - 1$ failures followed by a success) was discarded and the replicate redrawn.

https://doi.org/10.1017/S1138741600001694 Published online by Cambridge University Press

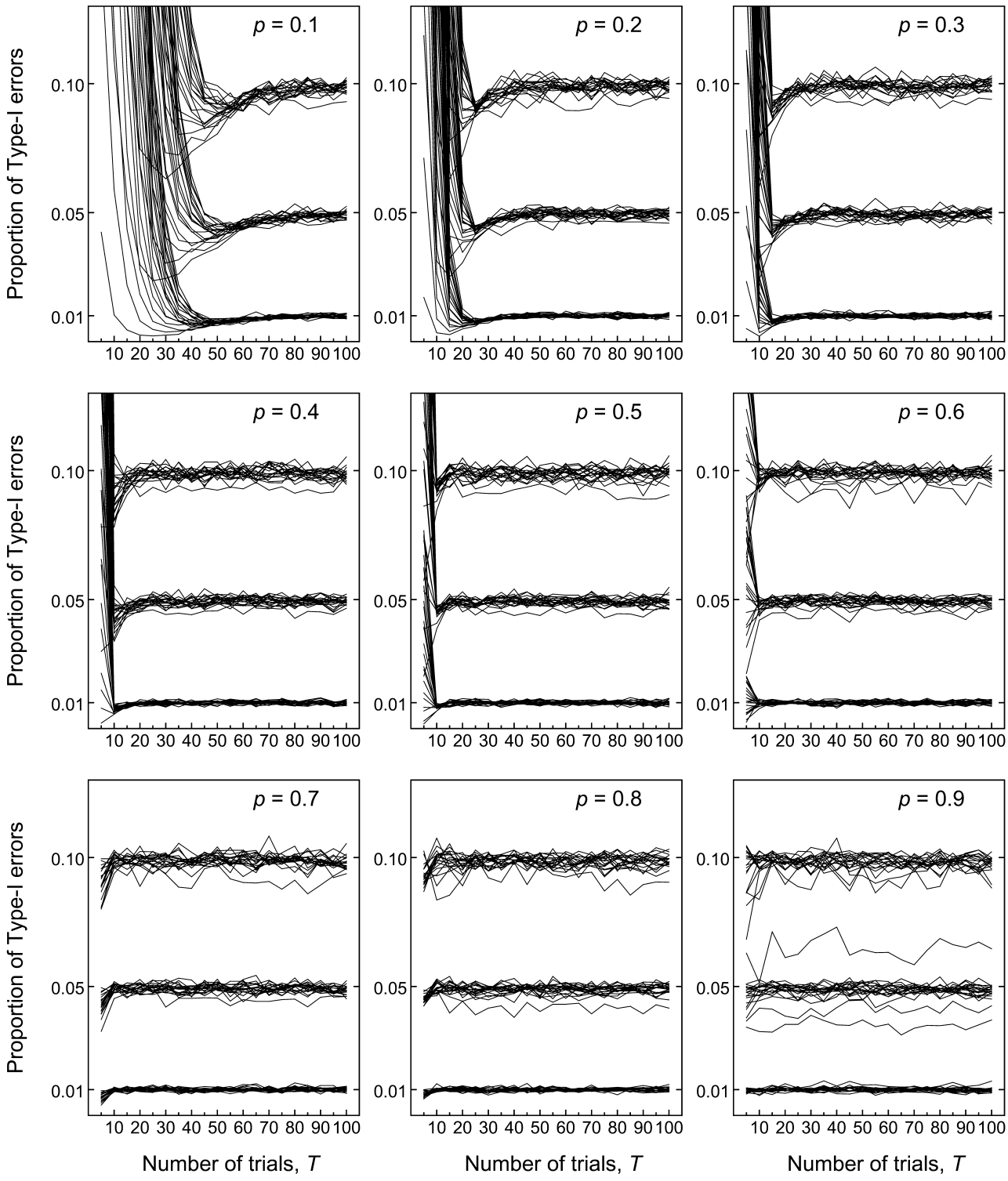


*Figure 3.* Empirical Type-I error rates in Case 1. Each panel shows results for a different true value of $p$. Tick marks on the vertical axis are drawn at the nominal test sizes $\alpha = .01$, .05, and .1. Data are plotted as a function of the number $T$ of trials; the number $R$ of replicates is the parameter that renders the 20 curves (unmarked) that thread each of the three bundles that respectively converge on the three nominal error rates as $T$ increases.

https://doi.org/10.1017/S1138741600001694 Published online by Cambridge University Press

Figure 2 are packed when the number $T$ of trials (along the abscissa in Figure 3) is sufficiently large, although how large must $T$ be for this to occur depends on the value of $p$. Nevertheless, differences in accuracy as a function of $R$ can be observed at a finer scale when both $p$ and $T$ are sufficiently large (see, e.g., the panel for $p = .9$, where some of the curves in each bundle appear to meander around the wrong ordinate), but this characteristic cannot be clearly seen in Figure 3 and will be described and illustrated in Figure 5 below.

The test is accurate even with $T = 10$ trials provided $p \geq .5$; otherwise, the test becomes increasingly liberal as $p$ decreases, and $T$ must hence be increasingly larger to restore accuracy. The reason for this inaccuracy is easy to understand on consideration of the characteristics of the data array, as illustrated in Figure 1. Small $T$ implies few trials per replicate and, if $p$ is also small, most of those trials are spent during the first phase of geometric sampling, leaving very few trials for the second phase of binomial sampling. Whether or not the number $R$ of replicates is large, $\hat{p}_b$ ends up being estimated from small $n_b$, violating the normality assumption that underlies the derivation of the test statistic (i.e., $n_b$ is too small for the central limit theorem to apply). The foregoing analysis suggests that a similar pattern of inaccuracy should occur when the number $R$ of replicates is small and $p$ is large: In this case, the first phase of geometric sampling would finish very early in each replicate and $\hat{p}_a$ would now be estimated from small $n_a$. Our results confirm this point, although Figure 3 cannot show this characteristic very clearly except for the fact that each of the three bundles of curves loosen out in the panels for large $p$ as compared to their pattern for intermediate $p$, revealing differences in accuracy according to the value of $R$ when $p$ is large.

From a practical point of view, some indication of the minimal values that $T$ and $R$ must have to guarantee accurate tests for each true $p$ would be useful. This, in turn, calls for a quantitative criterion that defines when a test is sufficiently accurate. This is an issue that seems to have received little attention in the literature, as most studies on the accuracy

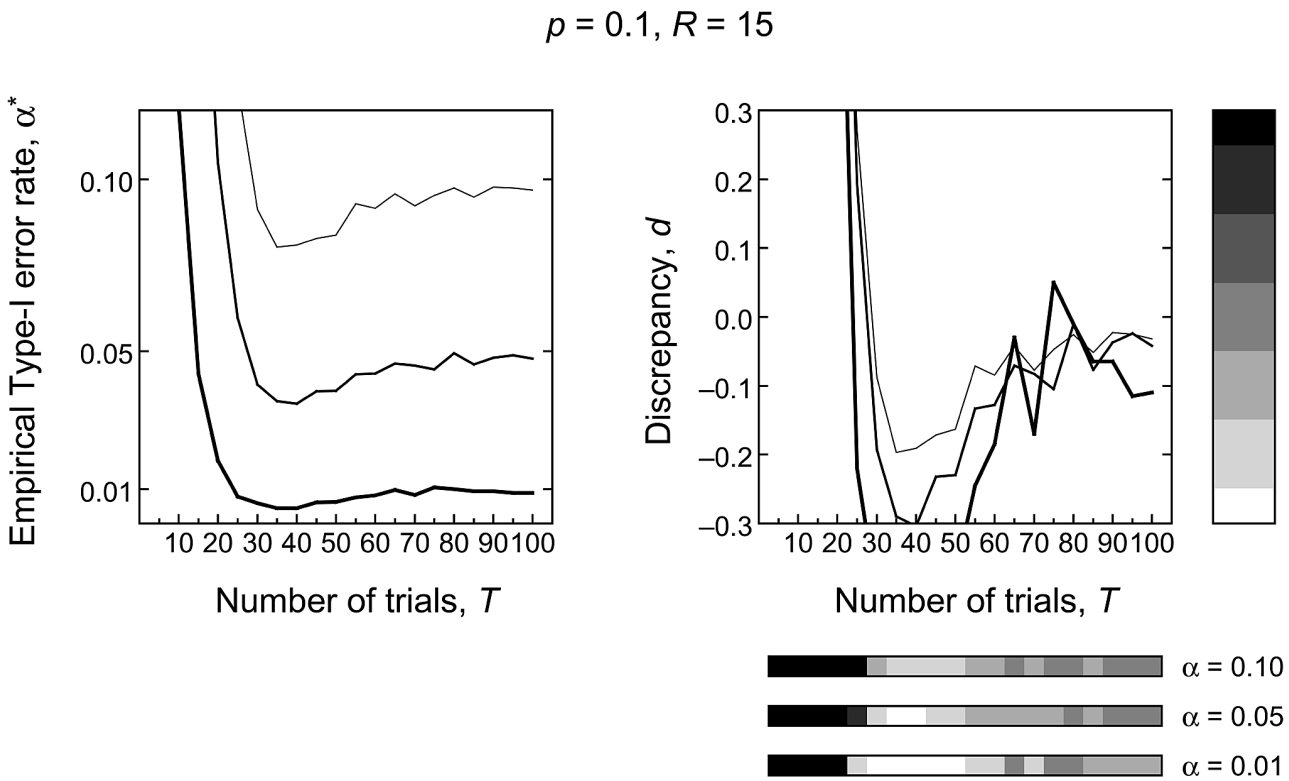


*Figure 4.* Left panel: Empirical Type-I error rates in Case 1 when $p = .1$ and $R = 15$, as a function of the number $T$ of trials (abscissa). The ordinate thus represents the empirical small-sample error rate $\alpha^*$ of a two-tailed test of nominal size $\alpha$, with $\alpha = .1$ (thin curve), .05 (mid-weight curve), and .01 (thick curve). Right panel: Discrepancy index obtained from each curve in the left panel by computing $d = (\alpha^* - \alpha)/\alpha$. Superposition of the three curves indicates that the relative inaccuracy of the test is similar at all nominal $\alpha$. The gray scale on the right indicates how the magnitude of $d$ is coded for subsequent display: Mid gray corresponds to accurate tests in which $-0.05 \leq d \leq 0.05$, whereas increasing darkness corresponds to increasingly liberal tests and increasing lightness corresponds to increasingly conservative tests as determined by the ranges of $d$ implied along the gray scale. All $d > 0.25$ is represented with black, and all $d < -0.25$ is represented with white. The three bars underneath are the gray-coded representations of the three curves shown in the panel.

https://doi.org/10.1017/S1138741600001694 Published online by Cambridge University Press

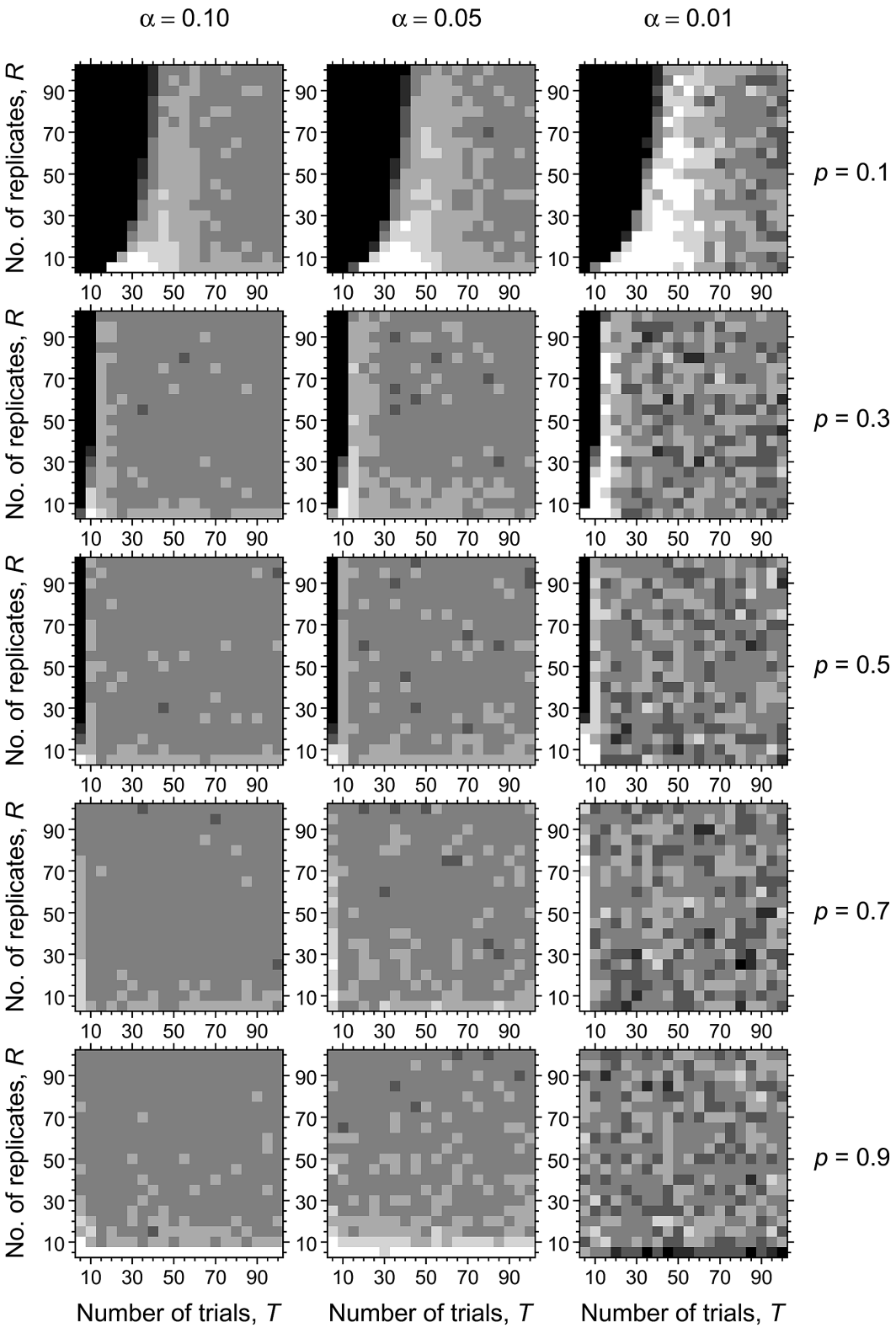


*Figure 5.* Small-sample accuracy of two-tailed tests in Case 1 with α = .1 (left column), .05 (center column), and .01 (right column) for selected values of $p$ (rows). The image in each panel shows accuracy according to the gray-scale code illustrated in Figure 4, as a function of the number $T$ of trials (abscissa) and the number $R$ of replicates (ordinate). Each image consists of a 20×20 array of blocks each of which pertains to the particular $T$ and $R$ at the coordinates of the block. Each of the three bars shown underneath the right panel in Figure 4 is located here as the third-from-bottom row of blocks (for $R$ = 15) in the appropriate column in the top row of images.

https://doi.org/10.1017/S1138741600001694 Published online by Cambridge University Press

of diverse tests have almost invariably reported only the raw values of empirical test sizes. In the analysis of contingency tables, Rudas (1986) addressed the issue of small-sample accuracy by determining whether the confidence interval around empirical critical limits contained the asymptotic critical limit, but this approach is unfeasible in our case because critical limits are table-dependent as described above.[2] We will use instead an index of discrepancy that was introduced by García-Pérez and Núñez-Antón (2004), namely, $d = (\alpha^* - \alpha)/\alpha$, where $\alpha^*$ is the empirical test size and $\alpha$ is the nominal test size. The discrepancy index $d$ quantifies the mismatch $\alpha^* - \alpha$ between actual and nominal test sizes by expressing it as a fraction of the nominal size $\alpha$ which is, then, interpretable as a proportion. Thus, a nominal size-.05 test actually yielding $\alpha^* = .04$ is regarded as equally inaccurate as a nominal size-.1 test actually yielding $\alpha^* = .08$: Both of them result in $d = -0.2$, implying that the test will reject a true null hypothesis 20% less often than expected from the nominal size $\alpha$. The sign of the index indicates whether the test is conservative (i.e., $\alpha^* < \alpha$, yielding negative $d$) or liberal (i.e., $\alpha^* > \alpha$, yielding positive $d$). The behavior of this index is illustrated in Figure 4 for two-tailed tests with $\alpha \in \{.1, .05, .01\}$ when $p = .1$ and $R = 15$ and for all values of $T$, where a gray-scale code is also introduced for subsequent use in our reporting of discrepancies. With this gray-scale code, mid gray corresponds to accurate tests for which $-0.25 \leq d \leq 0.25$ whereas increasing darkness corresponds to increasingly liberal tests and increasing lightness corresponds to increasingly conservative tests with the ranges of $d$ indicated by the key bar on the right of Figure 4. In addition, all $d > 0.25$ are represented in black and all $d < -0.25$ are represented in white.

Figure 5 uses the gray scale just introduced to report two-tailed size discrepancy at each pairing of $R$ and $T$ for selected values of $p$. Note that for $p = .1$ (topmost row of images) the test is either two liberal (black areas, where $d > 0.25$) or too conservative (white areas, where $d < -0.25$) for all $\alpha$ when $T \leq 40$ trials regardless of the number $R$ of replicates; conversely, for $p = .9$ (bottommost row of images) the test is also either too liberal or too conservative for all $\alpha$ when $R \leq 10$ replicates regardless of the number $T$ of trials per replicate. It should be noted that index $d$ does not carry with it any decision as to when a test is unacceptably liberal or conservative. The gray-scale code with which our data are presented in Figure 5 reveals our endorsing the position that departures larger than ±25% from the nominal $\alpha$ are unacceptable and, thus, conditions represented with black or white blocks in the images of Figure 4 should be avoided. This criterion is a little more permissive than the limit of ±10% used by García-Pérez and Núñez-Antón (2004) or the limit of ±20% used by Serlin and Harwell (2004) but not so much as the limit of ±40% implied by Larntz (1978) when he claimed that empirical levels within .02 of his nominal $\alpha = .05$ were acceptable. Nevertheless, we admit that how permissive should one be must depend on the particular application and that, providing this allowance, there is still ample room for debate.

### *Small-Sample Accuracy in Case 2*

Because the test should be accurate in this case regardless of the true value of parameter $p$, a first simulation generated 50,000 tables each of which had a true $p$ drawn from a uniform distribution on [.1, .9]. Figure 6 summarizes empirical Type-I error rates in this case, with left-tail and right-tail rates separately displayed. Overall and provided $T \geq 30$, the test is minimally liberal at nominal levels of .025 and higher, and it is sufficiently accurate when $\alpha \leq .01$. Note also that the patterns of accuracy (or lack thereof) are similar in both tails, that is, in the upper and lower halves of Figure 6. This feature reveals that one-tailed tests are equally accurate or inaccurate regardless of which tail is involved and that the difference between actual and nominal sizes for two-tailed tests doubles the magnitude of the corresponding difference for one-tailed tests.

But it is still interesting to check whether the test behaves differently for different $p$ (which it should not). Thus, we ran a second set of simulations as in the preceding section, that is, with 20,000 tables at each true $p$ (with $p$ between .1 and .9 in steps of .1) from which error rates were separately determined at each true $p$. The results (not shown) yielded a pattern at each $p$ that was very similar to the overall pattern shown in Figure 6 but also sharing the major characteristics found in Case 1 and displayed in Figure 2: inaccuracy for small $T$ when $p$ is also small that turns into adequate behavior as $p$ increases.

## Small-Sample Power

To evaluate power, data sets were generated under the alternative hypothesis that $p_a = p_b$ with a different value from that stated in the null hypothesis in Case 1, and that $p_a > p_b$ or $p_a < p_b$, for different values of $p_b$ and $p_a$ in Case 2. Then, appropriate critical limits described above were used to assess rejection rates. Simulations ran for the same combinations of $T$ and $R$ as in the preceding section.

[2] Note that, even if critical limits were fixed, confidence intervals for estimates of test size based on samples of 20,000 would be infinitesimally small.

https://doi.org/10.1017/S1138741600001694 Published online by Cambridge University Press

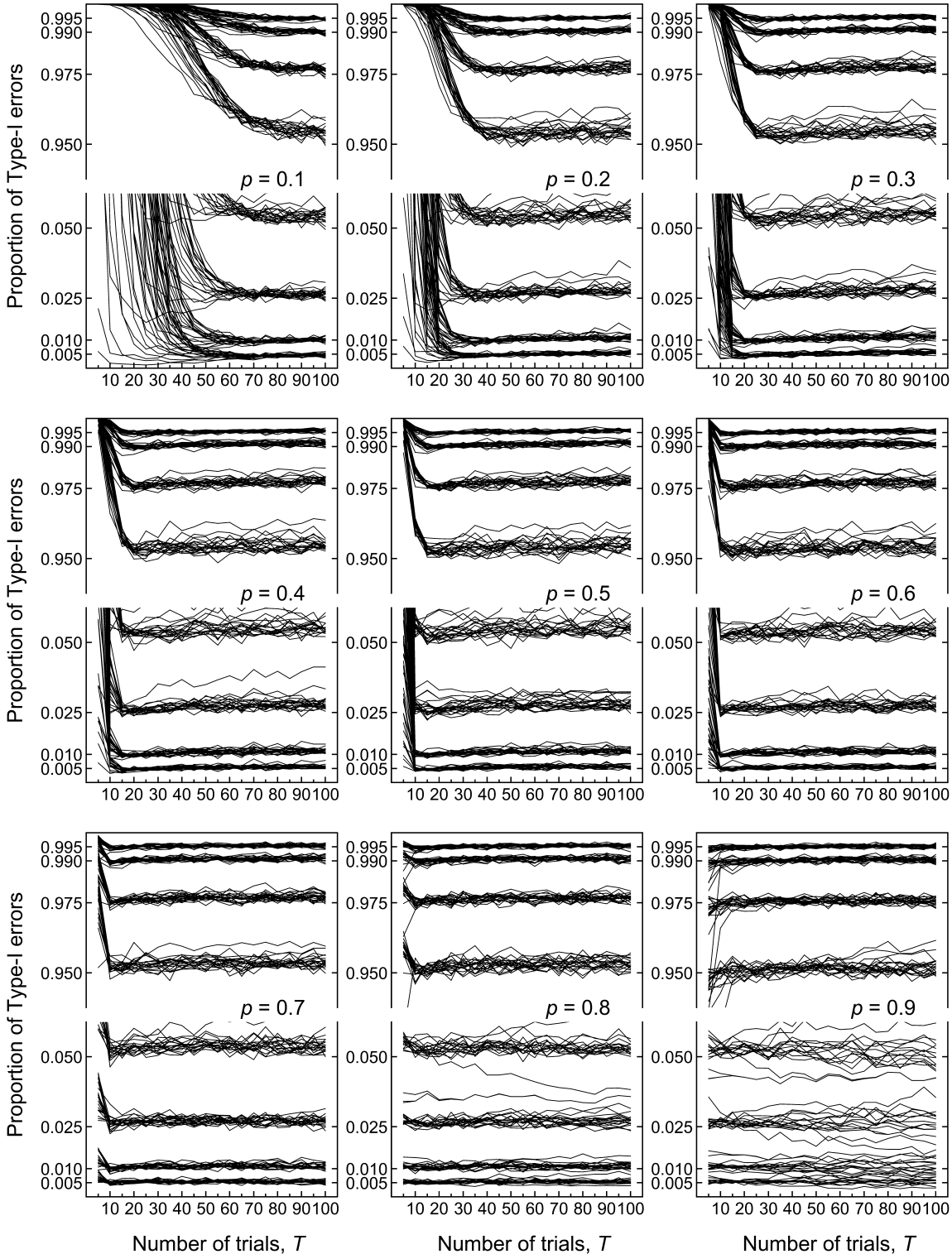


*Figure 6.* Empirical Type-I error rates in Case 2. Tick marks on the split vertical axis are drawn at the nominal probabilities that the statistic falls *below* each of the two critical points for a two-tailed size-α test with α = .01, .02, .05, and .1. To avoid clutter, empirical error rates are reported—quite unconventionally—by indicating the proportion of times that the statistic fell *below* each of the critical points that are designated by tick marks along the ordinate. Thus, data in the lower part actually indicate error rates, whereas data in the upper part indicate one's complement of error rates. Data are plotted as a function of the number $T$ of trials; the number $R$ of replicates is the parameter that renders the 20 curves (unmarked) that thread each of the eight bundles approximately converging on the nominal error rates as $T$ increases.

https://doi.org/10.1017/S1138741600001694 Published online by Cambridge University Press

*Small-Sample Power in Case 1*

Given the null hypothesis $H_0$: $p_a = p_b = p_0$, power was assessed by generating data for which $p_a = p_b = p_1$ with $p_1 \neq p_0$. We considered testing null hypotheses where $p_0$ varies from .1 to .9 in steps of .1 and, in each case, $p_1$ varied also from .1 to .9 in steps of .01. Figure 7 shows sample results for the cases $R = 60$, $T = 10$ (left column) and $R = 60$, $T = 25$ (right column) as a function of the value of $p_1$ when $p_0 = .3$ (top row), $p_0 = .5$ (center row), and $p_0 = .7$ (bottom row). Consider, for instance the case represented in the bottom row of Figure 7. When $\alpha = .1$ (topmost curve in each panel), testing $H_0$: $p_a = p_b = .7$ against $H_1$: $p_a = p_b = .75$ yields a power of .642 when $R = 60$ and $T = 10$ (left panel in the bottom row of Figure 7) and a power of .809 when $R = 60$ also but $T = 25$ (right panel in the bottom row). Adopting the common requirement that power be at least .8 (Clark-Carter, 1997; Cohen, 1992; Maxwell, 2004), the test is sufficiently powerful to detect a difference of .05 in probability when $R = 60$ but only if $T = 25$. In any case, inspection of Figure 7 reveals that the test has adequate power at all $p_0$ provided that $R$ and $T$ are sufficiently large for the conditions for small-sample accuracy to hold (i.e., in all the cases depicted in Figure 7 except when $R = 60$ and $T = 10$ with $p_0 = .3$; see the corresponding blocks in Figure 5). Note also that the test seems equally powerful to detect differences on either side except in non-accurate cases like the one just mentioned.

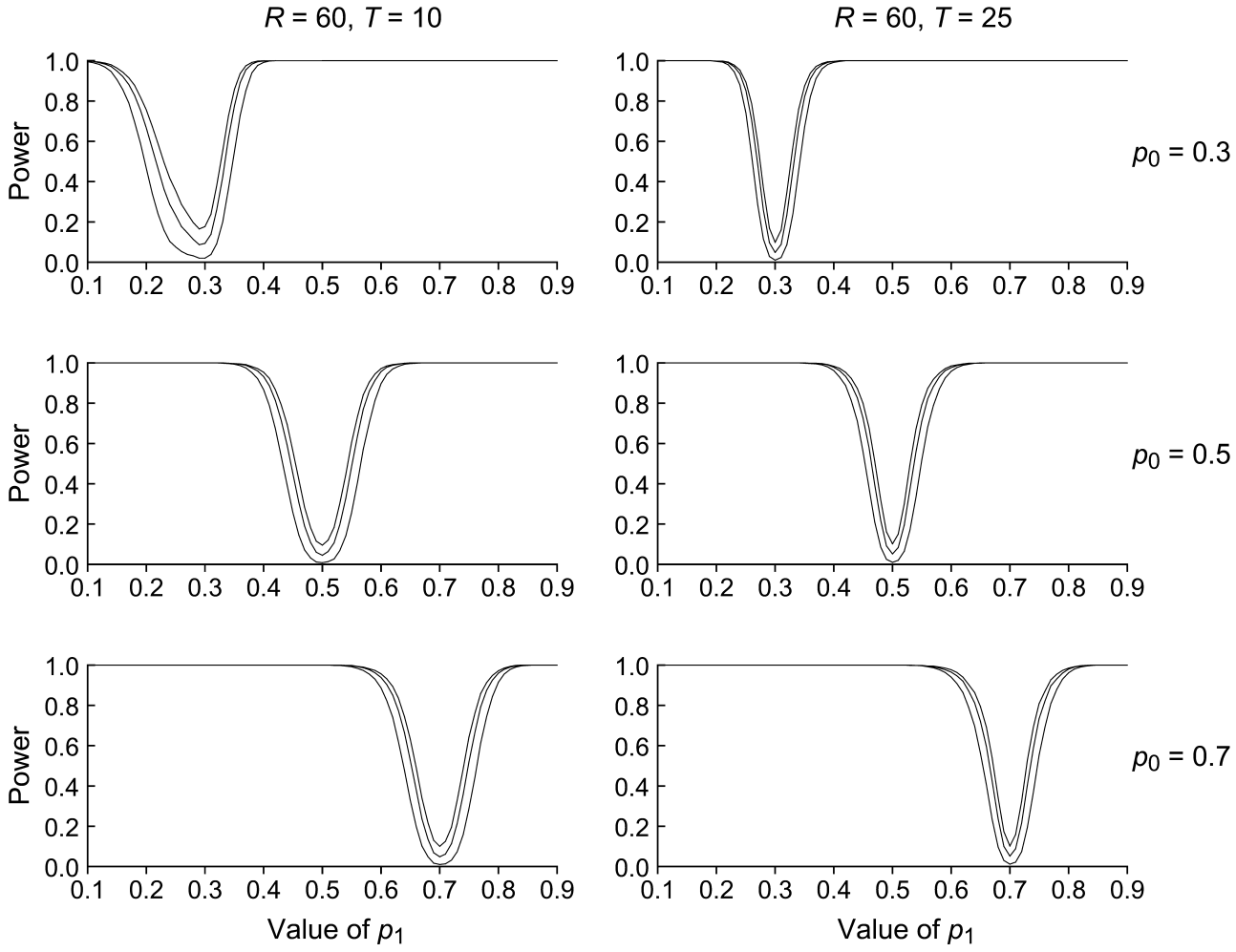


*Figure 7.* Sample results on power in Case 1 for two pairings of $R$ and $T$ (with values given at the top of each column) when testing the null hypothesis $H_0$: $p_a = p_b = p_0$ (with values for $p_0$ given on the right of each row) as a function of the actual value $p_1$ of the two binomial parameters in the population. The three curves in each panel pertain to three test sizes ($\alpha = .1$, .05, and .01). Note that the power function is similar—though appropriately displaced—at all $p_0$ when the test is accurate (i.e., in all conditions except the top-left panel; see the corresponding blocks in Figure 5) and that the spread of the power function shrinks as *RT* increases.

https://doi.org/10.1017/S1138741600001694 Published online by Cambridge University Press

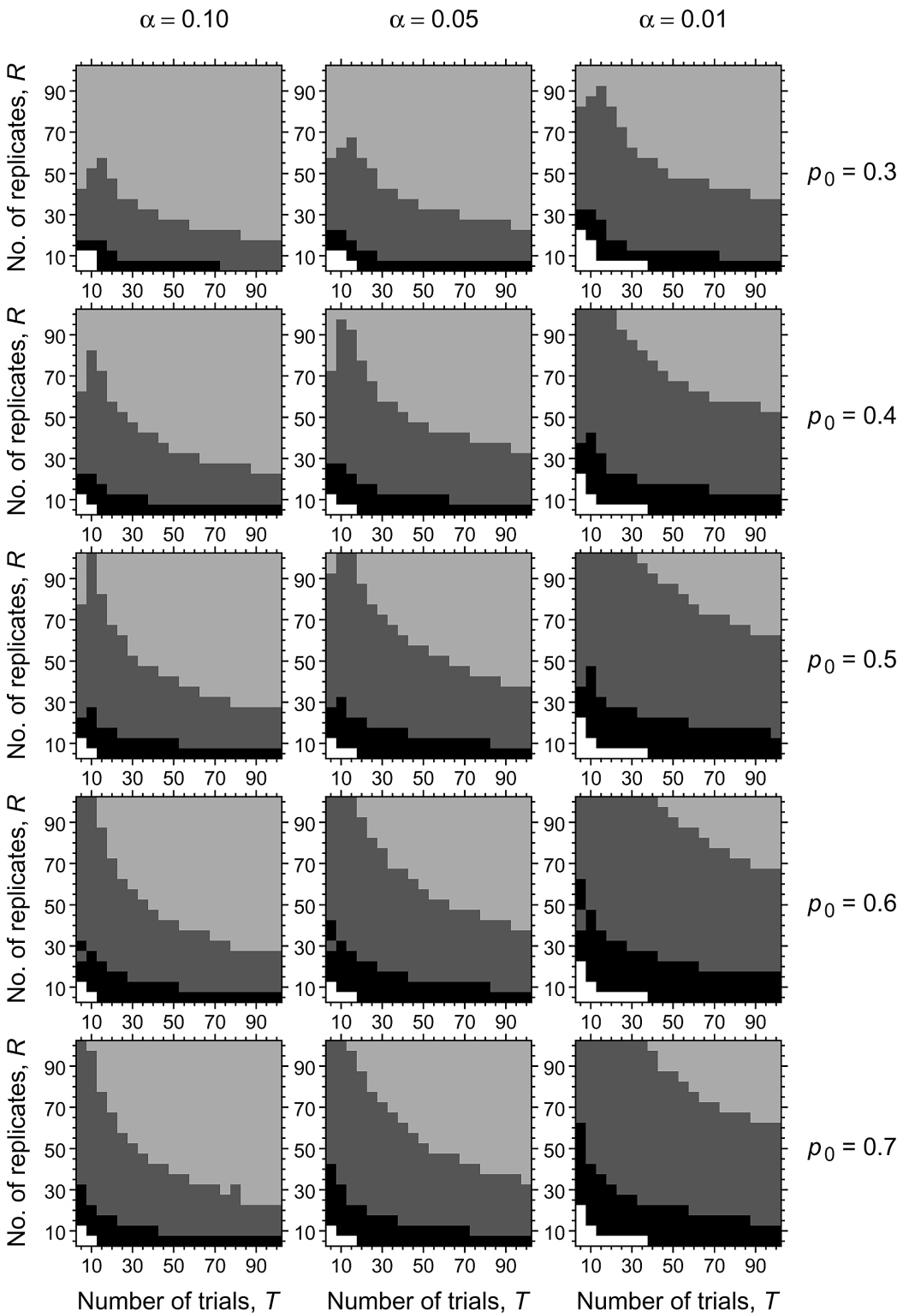


*Figure 8.* Small-sample power of two-tailed tests in Case 1 with α = .1 (left column), .05 (center column), and .01 (right column) for selected values of $p_0$ (rows). The image in each panel shows a gray code for each combination of the number $T$ of trials (abscissa) and the number $R$ of replicates (ordinate), where light gray, dark gray, and black respectively indicate that power is at least .8 to detect a small, medium, or large effect, defined in turn as $|p_1 - p_0|$ = .05, .1, and .2. The test may not be equally powerful for values of $p_1$ at the same distance from $p_0$ but on opposite sides (as illustrated in the top-left panel of Figure 7), but here a block of a given shade of gray indicates that the corresponding effect can actually be detected on either side.

https://doi.org/10.1017/S1138741600001694 Published online by Cambridge University Press

For a more complete picture, we determined sample size (i.e., $T$ and $R$) requirements for power to be at least .8 for detecting differences $|p_1 - p_0| = .05$, .10, or .20 (which we will loosely refer to as small, medium, and large effects) as a function of the null value $p_0$ (between .3 and .7) and at each of three test sizes ($\alpha = .1$, .05, and .01). Figure 8 shows the results, where the smallest effect that can be detected for each combination of $T$ (abscissa in each panel) and $R$ (ordinate in each panel) is indicated by a shaded block at the corresponding coordinates: light gray, dark gray, and black respectively indicate the possibility of detecting a small, a medium, and a large effect, whereas white blocks indicate that there is not enough power to detect even a large effect with the given choices for $T$ and $R$. Clearly, the capability of the test to detect small effects increases with $R$ and $T$, and the test appears slightly more powerful at intermediate $p_0$.

One other situation in which power should be evaluated is in the case that, contrary to what the null hypothesis assumes, $p_b \neq p_a$. When this is the case, one would expect the test statistic to reject the null hypothesis quite frequently, and results presented in Figure 9 show that this is actually the case: Rejection rates are quite high as soon as either $p_b$ or $p_a$ are minimally different from $p_0$.

### *Small-Sample Power in Case 2*

Given the null hypothesis $H_0$: $p_b/p_a = 1$, power was assessed by generating data for which $p_a \neq p_b$ and, given that the test can be left or right tailed, the power of each type of test was assessed separately. Values for $p_a$ varied between .1 to .9 in steps of .1 and, in each case, $p_b$ varied also from .1 to .9 in steps of .01. Figure 10 shows sample results for the case $R = 30$, $T = 40$ as a function of the value

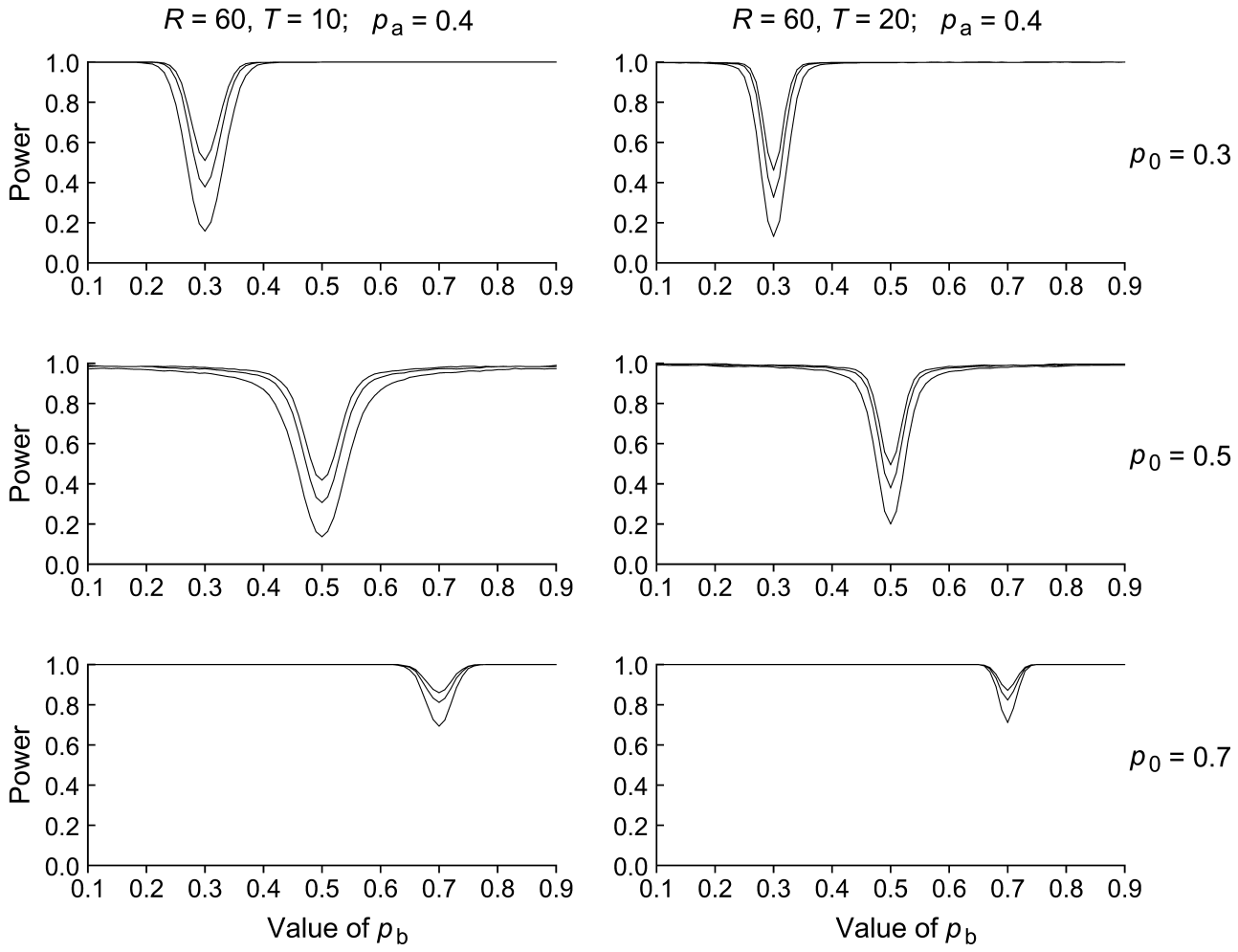


*Figure 9.* Sample results on power in Case 1 for the same conditions depicted in Figure 7 except that now $p_a = .4$ (and, then, generally $p_a \neq p_b$ in contrast with what the null hypothesis assumes).

https://doi.org/10.1017/S1138741600001694 Published online by Cambridge University Press

of $p_b$ when $p_a = .3$ (top row), $p_a = .5$ (center row), and $p_a = .7$ (bottom row); the left column depicts power against $H_0$: $p_b/p_a < 1$ and the right column depicts power against $H_0$: $p_b/p_a > 1$. Again, in all cases the test has sufficient and similar power to test against one-sided alternatives on either side, regardless of the slight inaccuracies that were reported above. Also, the test appears less powerful than in Case 1, considering that the power functions in Figure 10 are broader than those in Figure 7 (for Case 1) despite the fact that the product $RT$ is larger here.

Sample size requirements for power were also determined as in the preceding section, but small, medium, and large effects were now respectively defined as $|p_b - p_a| = .15$, .20, or .25. Results are shown in Figure 11, which were obtained for $p_a$ between .3 and .7 and for left and right tail tests of two sizes ($\alpha = .05$ and .01). The most salient characteristic of these results is that power increases with $R$ but is largely unaffected by $T$.

## Robustness to Violations of the Assumptions

In all of the simulations reported thus far parameter $p$ did not vary across replicates. In actual practice, each replicate may involve a different experimental subject or otherwise non-identical conditions. This characteristic will introduce variations in $p$ across replicates and, then, the issue arises as to whether the tests are robust to these variations. To assess robustness, a simulation study was carried out that was thoroughly analogous to the ones used to evaluate accuracy of the tests, except that now $p$ was not constant but varied randomly across replicates. For

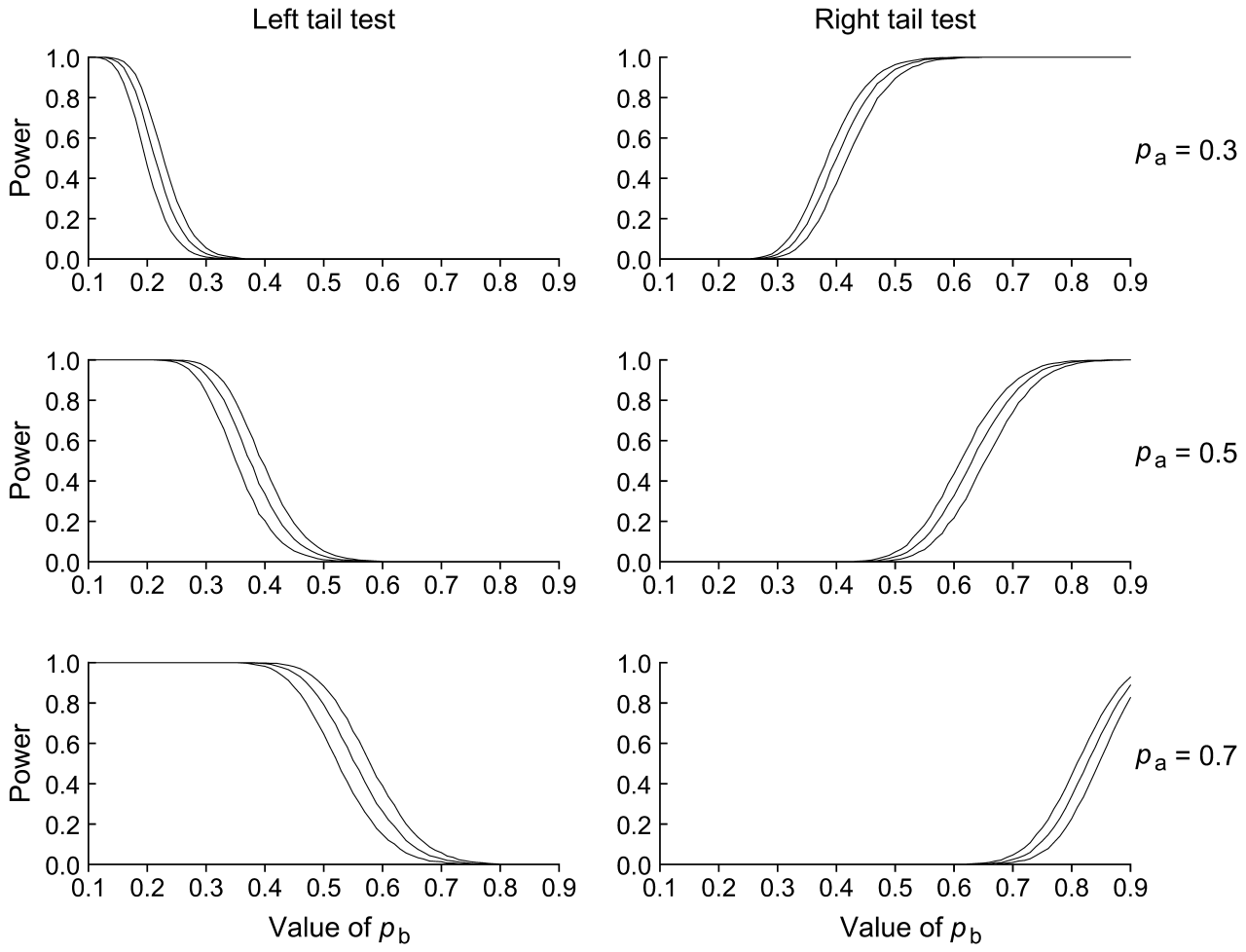


*Figure 10.* Sample results on power in Case 2 for $R = 30$ and $T = 40$ when testing $H_0$: $p_b/p_a = 1$ (with values for $p_a$ given on the right of each row) as a function of the actual value of $p_b$. The three curves in each panel pertain to three test sizes ($\alpha = .05$, .025, and .01). The left column shows the power of a left tail test (i.e., a test against $H_1$: $p_b/p_a < 1$) and the right column shows the power of a right tail test (i.e., a test against $H_1$: $p_b/p_a > 1$).

https://doi.org/10.1017/S1138741600001694 Published online by Cambridge University Press

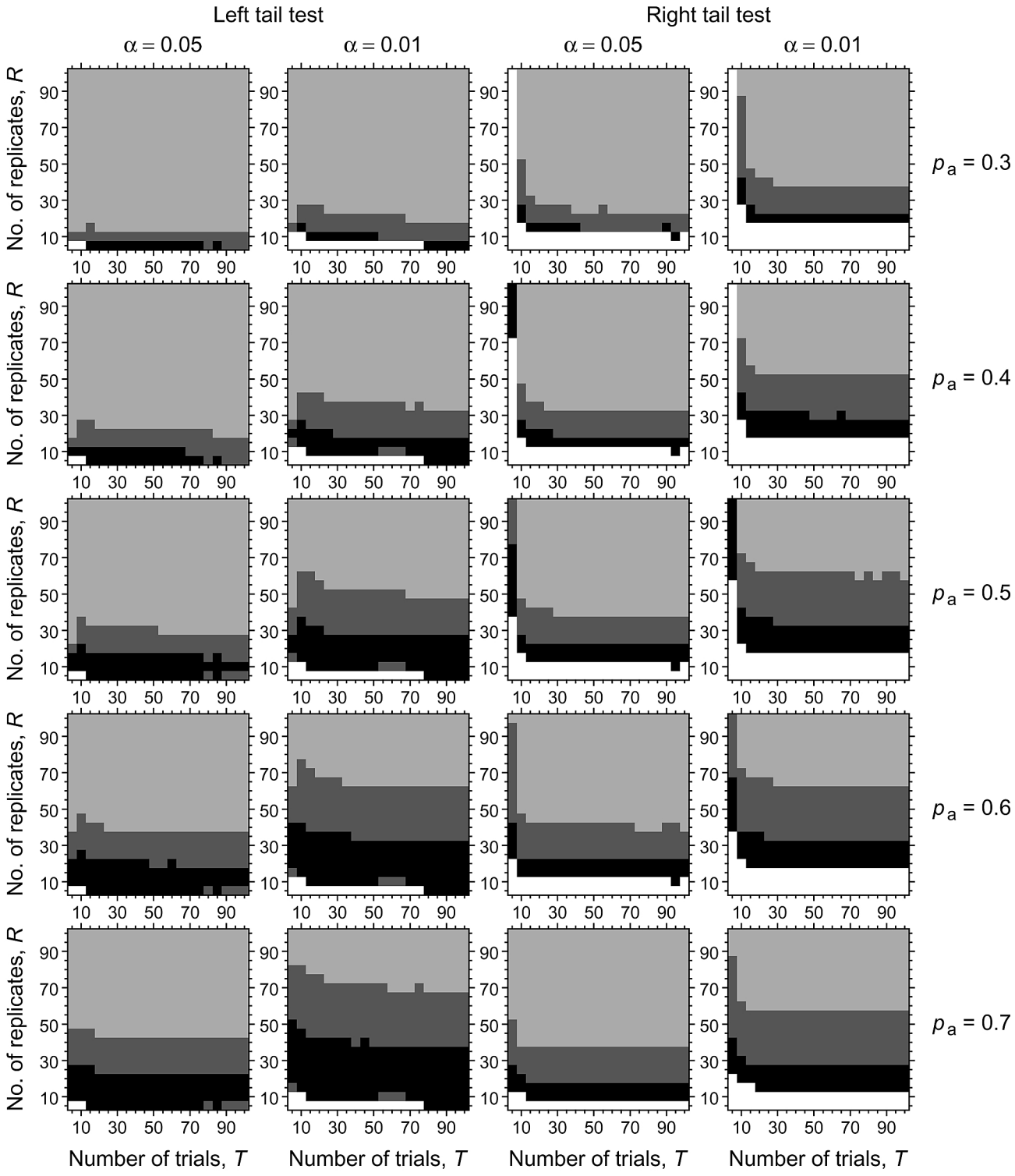


*Figure 11.* Small-sample power of left-tailed tests (left pair of columns) and right-tailed tests (right pair of columns) in Case 2 with α = .05 (left column within each pair) and .01 (right column within each pair) for selected values of $p_a$ (rows). Gray codes as in Figure 8, but light gray, dark gray, and black respectively indicate now small, medium, and large effects of size $|p_b - p_a|$ = .15, .20, and .25.

https://doi.org/10.1017/S1138741600001694 Published online by Cambridge University Press

this purpose we followed a strategy introduced by Riefer and Batchelder (1991) and, thus, when the nominal probability of success was $p$, the probability $p_r$ in replicate $r$ (with $1 \leq r \leq R$) was drawn from a beta distribution with parameters

$$v = \frac{p\,(p - p^2 - \sigma^2)}{\sigma^2}, \tag{13}$$

$$w = \frac{(1 - p)\,(p - p^2 - \sigma^2)}{\sigma^2}, \tag{14}$$

so that the mean is actually $p$ and the standard deviation is $\sigma$, which varied between 0.01 and 0.05 across simulations.

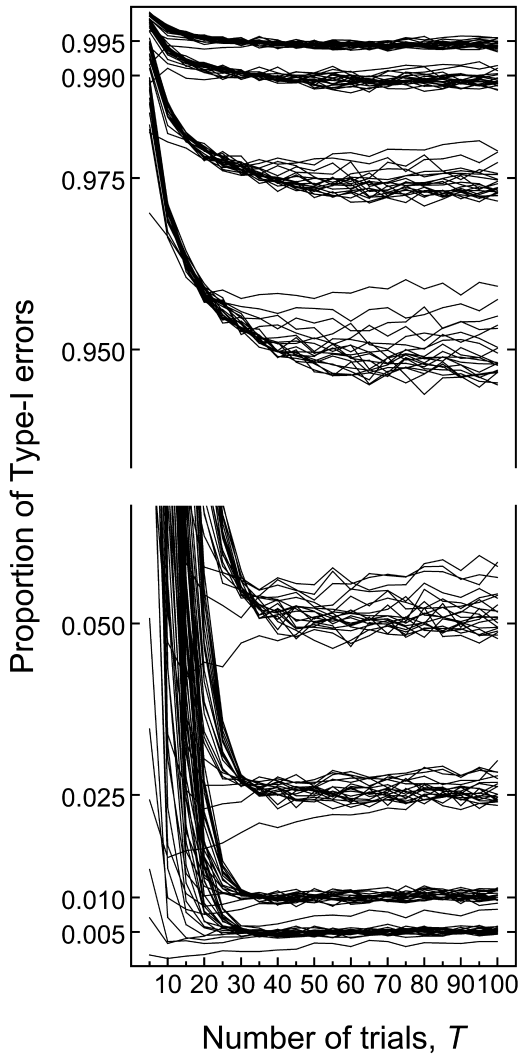


*Figure 12.* Empirical Type-I error rates in Case 2 under parameter heterogeneity. Simulation conditions are identical to those which produced the results in Figure 6 except that each of the $R$ replicates (which had the same, uniformly-distributed random $p$ in Figure 6) has now a different true $p$ arising from a beta distribution with mean $p$ (uniformly-distributed again) and standard deviation 0.02. All graphical conventions as in Figure 6.

In Case 1, parameter heterogeneity produced an increase in Type-I error rates when $p$ was close to its boundaries, particularly when $R$ was small. On the other hand, error rates remained fairly accurate at intermediate $p$ when $\sigma \leq 0.02$. This deterioration is not surprising because Case 1 tests for a particular value of $p$ which, in actual fact, does not exist as a constant across the $R$ replicates. Moreover, the average $p$ across a small number $R$ of replicates (which might be thought of as the "true" $p$ in these circumstances) rarely matches the nominal $p$ in the null hypothesis, particularly when $R$ and the nominal $p$ are both small and the beta distribution is very skewed. Thus, the null hypothesis is always bound to be literally false even on average, which accounts for the larger Type-I error rates. In Case 2, on the other hand, the test was naturally much more robust and heterogeneity of true $p$ only had the effect (shown in Figure 12) of producing occasionally more accurate Type-I error rates than in the absence of parameter heterogeneity (compare with Figure 6).

One other situation that may occur in practice is that the probability of success continues to vary as the number of successes increases, in contrast with the assumption thus far that it changes only after the first success. In these cases, the probability of success will generally vary monotonically as opposed to cyclically (i.e., it will either continue to increase or continue to decrease with the number of successes). The net effect will then be that the average probability of success in the second stage of binomial sampling will be much more different from that during the first stage of negative binomial sampling. We will not present simulation results in these conditions because it is clear from power results presented above that the methods developed in this paper will detect this situation, although these methods are certainly inadequate to test for specific trends in the pattern of change of the binomial parameter.

## Example

Apfelbaum, Apfelbaum, Woods, and Peli (2008) used the "selective looking" experiment of Neisser and Becklen (1975) to study the potential benefits of a vision multiplexing device to aid the visually impaired. The subjects had to attend to one of two scenes that were viewed simultaneously and perform some counting task while unexpected events might be occurring in the unattended scene. Subjects were asked at the end whether they had seen any of the unexpected events, and Apfelbaum et al. expected some priming in the form that subjects could have a higher chance of detecting unexpected events after they had detected the first such event. These are exactly the conditions for the methods derived in this paper. Thirty-six subjects confronted six trials in which unexpected events occurred in the

https://doi.org/10.1017/S1138741600001694 Published online by Cambridge University Press

unattended scene. One of the subjects did not detect any of the events and, thus, the string of six zeroes from this subject was discarded. Hence, with 35 subjects (replicates) $R = 35$ and with six trials per subject $T = 6$ for a $35 \times 6$ data array involving 210 total trials for analysis. As it turns out, there were 69 trials in the first sequence, with $x_a = R = 35$ successes (one per subject) and $y_a = 34$ failures for an estimate $\hat{p}_a = 35/69 = .5072$. This left $n_b = 141$ trials for the second sequence of which $x_b = 88$ were successes across subjects, yielding $\hat{p}_b = 88/141 = .6241$. Thus, $B' = \hat{p}_b / \hat{p}_a = .6421/.5072 = 1.230$. The probability and distribution functions that apply in this case were shown in the right column of Figure 2. At $\alpha = .05$, the critical limit for a right-tail test with $R = 35$, $T = 6$, and $n_b = 141$ is $b'_{.95} = 1.224$ (computed with the FORTRAN program referred to earlier in this paper). The $p$-value is indeed 0.0468. Hence, the data of Apfelbaum et al. reveal the presence of priming in that the probability of detecting unexpected events in the unattended task is significantly higher after the first such event has been detected.

## Conclusion

We have derived a method to test hypotheses involving the equality of two binomial parameters which are respectively estimated by negative binomial and by binomial sampling in related samples. Two different cases have been considered: Case 1 states that the two binomial parameters equal some specified value and yields a two-sided test, whereas Case 2 states that the two parameters are simply equal to one another and yields one-sided or two-sided tests. Simulations have shown that in both cases the test is sufficiently accurate and powerful in small samples, and in Case 2 the test is also reasonably robust to violations of the assumption that the binomial parameter does not vary across replicates.

We have not addressed confidence intervals for different reasons in each of the two cases. In Case 1, a closed-form expression for the confidence interval of the common binomial parameter under the null hypothesis cannot be derived, but the confidence interval can easily be obtained numerically by solving for $p$ the well-known inequality $b_{\alpha/2} < \frac{\hat{p}_b - p}{\sqrt{p\,(1-p)/n_b}} \; \frac{1/\hat{p}_a - 1/p}{\sqrt{(1-p)/R\,p^2}} < b_{1-\alpha/2}$, where $b_{\alpha/2}$ and $b_{1-\alpha/2}$ are quantiles from the Bessel distribution given in Appendix B. In Case 2, on the other hand, a confidence interval cannot be derived for lack of the term $p_b/p_a$ in the corresponding test statistic.

## References


Apfelbaum, H. L., Apfelbaum, D. H., Woods, R. L., & Peli, E. (2008). Inattentional blindness and augmented-vision displays: Effects of cartoon-like filtering and attended scene. *Ophthalmic and Physiological Optics*, *28*, 204–217.

Bain, L. J., & Engelhardt, M. (1992). *Introduction to probability and mathematical statistics* (2nd edition). Pacific Grove, CA: Duxbury.

Clark-Carter, D. (1997). The account taken of statistical power in research published in the *British Journal of Psychology*. *British Journal of Psychology*, *88*, 71–83.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159.

García-Pérez, M. A., & Núñez-Antón, V. (2004). Small-sample comparisons for goodness-of-fit statistics in one-way multinomials with composite hypotheses. *Journal of Applied Statistics*, *31*, 161–181.

Larntz, K. (1978). Small-sample comparisons of exact levels for chi-squared goodness-of-fit statistics. *Journal of the American Statistical Association*, *73*, 253–263.

The MathWorks, Inc. (2004). *MATLAB: The language of technical computing*. Natick, MA: Author.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods*, *9*, 147–163.

Neisser, U., & Becklen, R. (1975). Selective looking: Attending to visually specified events. *Cognitive Psychology*, *7*, 480–494.

Numerical Algorithms Group (1999). *NAG Fortran library manual, Mark 19*. Oxford: Author.

Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. T. (1986). *Numerical recipes: The art of scientific computing*. New York: Cambridge University Press.

Riefer, D. M., & Batchelder, W. H. (1991). Statistical inference for multinomial processing tree models. In J.-P. Doignon & J.-C. Falmagne (Eds.), *Mathematical psychology: Current developments* (pp. 313–335). New York: Springer.

Rudas, T. (1986). A Monte Carlo comparison of the small sample behaviour of the Pearson, the likelihood ratio and the Cressie–Read statistics. *Journal of Statistical Computation and Simulation*, *24*, 107–120.

Serlin, R. C., & Harwell, M. R. (2004). More powerful tests of predictor subsets in regression analysis under nonnormality. *Psychological Methods*, *9*, 492–509.

Visual Numerics, Inc. (1997). *IMSL math/library special functions*. Houston, TX: Author.

Wolfram, S. (1992). *Mathematica: A system for doing mathematics by computer* (2nd edition). New York: Addison-Wesley.


Received April 11, 2008
Revision received July 24, 2008
Accepted September 21, 2008

# APPENDIX A

The probability distribution of a transformation of two random variables is obtained by marginalizing the product of their joint density evaluated at the variables expressed in terms of the transformation and the absolute value of the Jacobian of the transformation (Bain & Engelhardt, 1992, p. 206).

We know that, asymptotically, $Y_a = \dfrac{1/\hat{p}_a - 1/p_a}{\sqrt{(1-p_a)/R\,p_a^2}}$ and $Y_b = \dfrac{\hat{p}_b - p_b}{\sqrt{p_b(1-p_b)/n_b}}$ are both distributed N(0, 1) and

the joint distribution of $Y_a$ and $Y_b$ can be written as $f(Y_a, Y_b) = f_1(Y_b \mid Y_a) \times f_2(Y_a)$. We are now interested in finding the probability distribution of the product of $Y_a$ and $Y_b$, so let $W = Y_aY_b$ and $V = Y_a$. Then, $Y_b = W/V$, $Y_a = V$, and the Jacobian of the transformation is

$$J = \begin{vmatrix} \partial Y_a/\partial W & \partial Y_a/\partial V \\ \partial Y_b/\partial W & \partial Y_b/\partial V \end{vmatrix} = \begin{vmatrix} 0 & 1 \\ 1/V & -W/V^2 \end{vmatrix} = 1/V. \tag{A1}$$

The joint density of $V$ and $W$ is then given by $f^*(V, W) = f(Y_a = V, Y_b = W/V) / |V|$. Therefore,

$$f^*(V, W) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}V^2\right] \times \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}\frac{W^2}{V^2}\right] \times \frac{1}{|V|},\ V, W \in \mathbb{R}. \tag{A2}$$

Finally, the marginal distribution of $W = Y_aY_b$ is obtained by integrating $V$ out, that is,

$$f^*(W) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}V^2\right] \times \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}\frac{W^2}{V^2}\right] \times \frac{1}{|V|}\mathrm{dV} = \frac{K_0(|W|)}{\pi},\ W \in \mathbb{R}. \tag{A3}$$

where $K_0$ is the modified Bessel function of the second kind and zero order. The left panel of Figure 13 plots the probability density of the Bessel distribution in Equation (A3).

Although the integral of the Bessel distribution cannot be obtained in closed form because of its singularity at zero, the cumulative density function can nevertheless be easily obtained. First note that the Bessel distribution is even symmetric about zero so that the cumulative density is odd symmetric about zero. Thus, for $x > 0$, let

$$g(x) = \int_0^x f^*(b)\,\mathrm{d}b = \int_0^x \frac{K_0(b)}{\pi}\mathrm{d}b = \frac{x}{2}\left(K_0(x)\,L_{-1}(x) + K_1(x)\,L_0(x)\right), \tag{A4}$$

where $K_1$ is the modified Bessel function of the second kind and order 1 and $L_n$ is the modified Struve function of order $n$. Then, finally, for $b \in \mathbb{R}$,

$$F^*(b) = \begin{cases} \frac{1}{2} - g(-b) & \text{if } b < 0 \\ \frac{1}{2} & \text{if } b = 0 \\ \frac{1}{2} + g(b) & \text{if } b > 0 \end{cases} \tag{A5}$$

The right panel of Figure 13 plots the cumulative density in Equation (A5).

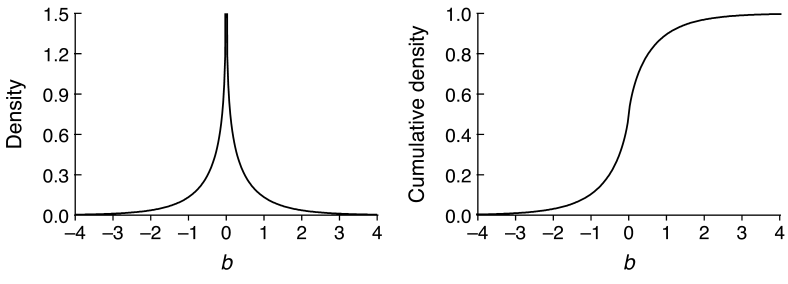


*Figure 13.* Probability density (left panel) and cumulative density (right panel) of a Bessel distribution.

https://doi.org/10.1017/S1138741600001694 Published online by Cambridge University Press

# APPENDIX B

The table below gives percentage points of the Bessel distribution. These are the points $b_\nu$ for which $P(B \leq b_\nu) = \nu$, where $B$ is a Bessel-distributed variable. The first two digits of $\nu$ are given at the left, and the third digit is given at the top. The Bessel distribution is even symmetric around zero and, therefore, $b_{1-\nu} = -b_\nu$. Thus, for a two-tailed size-.05 test, the critical limits are $b_{.975} = 2.18195$ and $b_{.025} = -2.18195$.

| $\nu$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| .90 | 1.03438 | 1.04226 | 1.05023 | 1.05829 | 1.06644 | 1.07469 | 1.08303 | 1.09147 | 1.10001 | 1.10866 |
| .91 | 1.11741 | 1.12627 | 1.13524 | 1.14432 | 1.15351 | 1.16283 | 1.17226 | 1.18182 | 1.19150 | 1.20132 |
| .92 | 1.21127 | 1.22135 | 1.23158 | 1.24195 | 1.25246 | 1.26313 | 1.27395 | 1.28494 | 1.29609 | 1.30740 |
| .93 | 1.31889 | 1.33056 | 1.34242 | 1.35446 | 1.36670 | 1.37914 | 1.39179 | 1.40465 | 1.41774 | 1.43105 |
| .94 | 1.44460 | 1.45840 | 1.47244 | 1.48675 | 1.50133 | 1.51619 | 1.53135 | 1.54680 | 1.56257 | 1.57867 |
| .95 | 1.59510 | 1.61189 | 1.62905 | 1.64659 | 1.66453 | 1.68289 | 1.70169 | 1.72095 | 1.74068 | 1.76092 |
| .96 | 1.78169 | 1.80301 | 1.82492 | 1.84744 | 1.87061 | 1.89447 | 1.91906 | 1.94442 | 1.97059 | 1.99764 |
| .97 | 2.02562 | 2.05459 | 2.08463 | 2.11581 | 2.14822 | 2.18195 | 2.21712 | 2.25385 | 2.29228 | 2.33257 |
| .98 | 2.37490 | 2.41948 | 2.46657 | 2.51645 | 2.56945 | 2.62600 | 2.68658 | 2.75179 | 2.82239 | 2.89933 |
| .99 | 2.98381 | 3.07745 | 3.18244 | 3.30182 | 3.44008 | 3.60421 | 3.80591 | 4.06715 | 4.43747 | 5.07546 |

https://doi.org/10.1017/S1138741600001694 Published online by Cambridge University Press

# APPENDIX C

The maximum likelihood estimate of a common parameter in two independent data sets is given by the value that maximizes the product of the individual likelihood of each data set. On the assumption that $p_a = p_b = p$, the maximum likelihood estimate $\hat{p}_\bullet$ of $p$ is, thus, the value that maximizes

$$L(x_a; p)\, L(x_b; p) = \begin{pmatrix} n_a - 1 \\ x_a - 1 \end{pmatrix} p^{x_a}(1-p)^{n_a - x_a} \begin{pmatrix} n_b \\ x_b \end{pmatrix} p^{x_b}(1-p)^{n_b - x_b} = \begin{pmatrix} n_a - 1 \\ x_a - 1 \end{pmatrix} \begin{pmatrix} n_b \\ x_b \end{pmatrix} p^{x_a + x_b}(1-p)^{n_a + n_b - x_a - x_b} \tag{C1}$$

Differentiating the logarithm of this expression with respect to $p$ and solving the resultant likelihood equation for $p$ yields

$$\hat{p}_\bullet = \frac{x_a + x_b}{n_a + n_b}, \tag{C2}$$

a result that can be easily understood from the similarity of Equation (C1) with the likelihood of pure binomial data.

https://doi.org/10.1017/S1138741600001694 Published online by Cambridge University Press